

(19) 世界知的所有権機関  
国際事務局(43) 国際公開日  
2004 年 7 月 15 日 (15.07.2004)

PCT

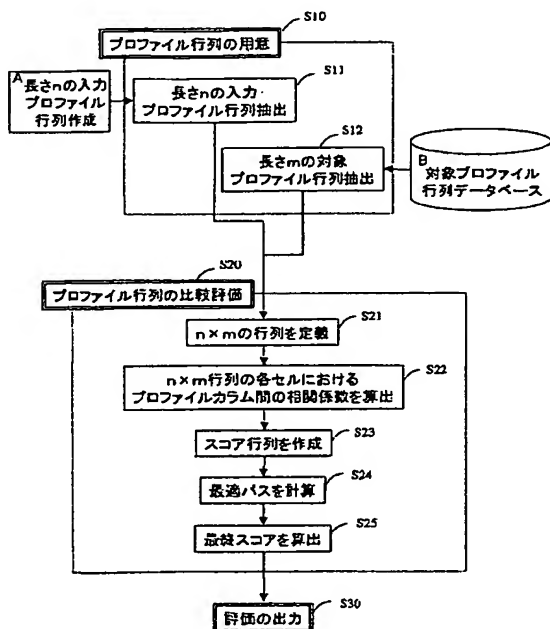
(10) 国際公開番号  
WO 2004/059557 A1

- (51) 国際特許分類: G06F 19/00 (71) 出願人 (米国を除く全ての指定国について): 独立行政法人産業技術総合研究所 (NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY) [JP/JP]; 〒100-8921 東京都千代田区霞が関 1 丁目 3 番 1 号 Tokyo (JP).
- (21) 国際出願番号: PCT/JP2003/016982
- (22) 国際出願日: 2003 年 12 月 26 日 (26.12.2003)
- (25) 国際出願の言語: 日本語 (72) 発明者; および (75) 発明者/出願人 (米国についてのみ): 富井 健太郎 (TOMII, Kentaro) [JP/JP]; 〒135-0064 東京都江東区青海 2-4 1-6 独立行政法人産業技術総合研究所内 Tokyo (JP).
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:  
特願 2002-377704 2002 年 12 月 26 日 (26.12.2002) JP  
特願 2003-406776 2003 年 12 月 5 日 (05.12.2003) JP (81) 指定国 (国内): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM,

(続葉有)

(54) Title: SYSTEM FOR PREDICTING THREE-DIMENSIONAL STRUCTURE OF PROTEIN

(54) 発明の名称: タンパク質立体構造予測システム



S10...PREPARE PROFILE MATRIX  
 A...PREPARE INPUT PROFILE MATRIX OF LENGTH n  
 S11...EXTRACT INPUT PROFILE MATRIX OF LENGTH n  
 S12...EXTRACT OBJECT PROFILE MATRIX OF LENGTH m  
 B...OBJECT PROFILE MATRIX DATABASE  
 S20...COMPARE AND EVALUATE PROFILE MATRIX  
 S21...DEFINE n x m MATRIX  
 S22...CALCULATE COEFFICIENT OF CORRELATION BETWEEN PROFILE COLUMNS IN EACH CELL OF n x m MATRIX  
 S23...MAKE SCORE MATRIX  
 S24...CALCULATE OPTIMUM PATH  
 S25...CALCULATE FINAL SCORE  
 S30...OUTPUT EVALUATION

(57) Abstract: A system for evaluating the similarity between protein profile matrices, preferably usable for prediction of the three-dimensional structure of a protein. A profile matrix is composed of profile columns provided with the appearance probabilities of amino acids at the positions of the amino acid residues in a multiple alignment in which the amino acid sequences of relevant proteins are multiply arranged. The similarity evaluating system comprises (a) means for preparing two matrices, an input profile matrix and an object profile matrix, (b) means for calculating the coefficient of correlation between a profile column of the input profile matrix and a profile column of the object profile matrix for all or a part of the combinations of the profile columns, and (c) means for making a score matrix composed of the coefficients of correlation.

(57) 要約: タンパク質の立体構造予測に好適に使用できる、タンパク質プロフィール行列間の類似性評価システムを提供する。本発明は、タンパク質プロフィール行列間の類似性を評価するシステムであって、プロフィール行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアライメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロフィールカラムの群から構成され、(a) 入力プロフィール行列と、対象プロフィール行列の2つのプロフィール行列を用意する手段と、(b) 前記入力プロフィール行列の各プロフィールカラムと、前記対象プロフィール行列の各プロフィールカラムとの間の相関係数を、各プロフィールカラムの全部又は一部の組合せについて算出する手段と、(c) 前記相関係数からなるスコア行列を作成する手段とを含む。



HR, HU, ID, IL, IN, IS, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI 特許 (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告書

(84) 指定国 (広域): ARIPO 特許 (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア特許 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ特許 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,

2 文字コード及び他の略語については、定期発行される各 PCT ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

## 明細書

## タンパク質立体構造予測システム

## 5 技術分野

本発明は、タンパク質プロファイル行列間の類似性を評価するシステムに関するものであり、より詳しくは、タンパク質の立体構造予測に好適に使用されるタンパク質プロファイル行列間の類似性の評価システムに関する。

## 10 背景技術

自然界にあるタンパク質は進化の過程で選択され、特定の機能を発現するに至ったが、このタンパク質の機能はその立体構造に依存することが知られている。したがって、タンパク質の立体構造が予測できれば、その機能を予測することが可能となる。

- 15 従来、未だ何の知見も得られていないタンパク質を調べるに際し、既に立体構造が知られているタンパク質との類似性をコンピュータによって測定することにより、タンパク質の立体構造を推論ないし予測する手法が望まれていた。このような手法の1つとして、タンパク質プロファイル行列同士を比較する方法が、有力な手法として知られている (Rychlewski L, Jaroszewski L, Li W, Godzik A. Protein Sci (2000) Feb;9(2):232-41: 非特許文献1)。

- 20 ここで、タンパク質プロファイル行列とは、関連するタンパク質 (タンパク質ファミリーなど) におけるアミノ酸種の出現頻度を、そのアミノ酸残基位置毎に数値化して行列としたものである。この行列は、通常、以下の手順で作成される。すなわち、まず、関連する複数のタンパク質のアミノ酸配列を多重並置させた
- 25 マルチプルアラインメントが与えられると、マルチプルアラインメントの各アミノ酸残基位置における20種のアミノ酸の各種類の出現数が計算される。続いて、これらの数を規格化することによって、出現確率に転換される。この時、与えられたマルチプルアラインメントに含まれるメンバー内での相互のアミノ酸配列類似性に応じた重みが考慮された上で出現数が補正され、プロファイル行列が作

成される。

ここで、マルチプルアラインメントとは、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列を、対応すると考えられるアミノ酸残基を揃えて並置したものをいう。マルチプルアラインメントは、例えば、ある一配列を入力値として、既存のプログラムであるPSI-BLAST(Altschul et al., Nucleic Acids Res. (1997) 25(17):3389-3402: 非特許文献2)を用いて、配列データベースに検索をかけることや、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列の一群を入力値として、これも既存のプログラムであるCLUSTALW(Higgins D., Thompson J., Gibson T. Thompson J.D., Higgins D.G., Gibson T.J. (1994). Nucleic Acids Res. 22:4673-4680: 非特許文献3)を用いることで容易に作成することができる。また、立体構造比較などの結果から作成することも可能である。

表1は、アミノ酸配列の長さ(アミノ酸残基数)がnであるタンパク質を基準として作成されたマルチプルアラインメントを模式的に示したものである。なお、表1中、第1列目は個々のタンパク質の名称であり、第1行目の「1~n」は、マルチプルアラインメントにおけるアミノ酸残基位置を示す。また、表1中のアルファベットはアミノ酸種を1文字標記したものである。

【表1】

	1	2	3	4	5	6	7	8	...	n
20807455/14-218	M	I	D	H	T	L	L	K	...	G
19551629/13-215	I	L	D	Y	T	L	L	G	...	A
16974933/15-229	L	M	D	L	T	T	L	N	...	A
16120769/20-234	L	M	D	L	T	T	L	N	...	A

表1の例では、例示されたアミノ酸残基位置のすべてにアミノ酸が配置されているが、アミノ酸残基位置に対応するアミノ酸残基がないとされた場合は、「・(ドット)」としてギャップを示すこともできる。表2は、表1で得られた長さがnであるマルチプルアラインメントにしたがって作成されたプロファイル行列を模式的に示したものである。表2中、第1列目はアミノ酸種(ギャップを含ん

でいてもよい) であり、第 1 行目の「1～n」は、プロファイル行列におけるアミノ酸残基位置を示す。

【表 2】

AA/Pos.	1	2	3	...	n
A	0.00	0.00	0.00	...	0.71
R	0.00	0.00	0.00	...	0.00
N	0.00	0.00	0.00	...	0.00
D	0.00	0.00	0.96	...	0.00
C	0.00	0.00	0.00	...	0.00
Q	0.00	0.00	0.00	...	0.00
E	0.00	0.00	0.04	...	0.00
G	0.00	0.00	0.00	...	0.29
H	0.00	0.00	0.00	...	0.00
I	0.29	0.29	0.00	...	0.00
L	0.41	0.29	0.00	...	0.00
K	0.00	0.00	0.00	...	0.00
M	0.29	0.41	0.00	...	0.00
F	0.00	0.00	0.00	...	0.00
P	0.00	0.00	0.00	...	0.00
S	0.00	0.00	0.00	...	0.00
T	0.00	0.00	0.00	...	0.00
W	0.00	0.00	0.00	...	0.00
Y	0.00	0.00	0.00	...	0.00
V	0.01	0.01	0.00	...	0.00

5

プロファイル行列中の各列は、関連する複数のタンパク質における、各アミノ酸残基位置の全アミノ酸種の確率分布を表すことになる。表 3 は、表 2 に示されたプロファイル行列のうち、残基位置が「2」であるプロファイルカラムを模式的に示したものである。

【表 3】

2
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.29
0.29
0.00
0.41
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.01

すなわち、表 2 で示されるプロファイル行列では、残基位置が 2 におけるアラニン (A) の補正された出現確率は 0.00 であり、メチオニン (M) の補正された出現確率は 0.41 ということになる。

従来、2つのプロファイル行列や2つのアミノ酸配列を比較及び／又は揃えるために、ダイナミックプログラミング (Needleman SB, Wunsch CD, J Mol Biol. (1970) Mar;48(3):443-53 : 非特許文献 4) が使用されてきた。アラインメントを作成する時に、比較される2つのアミノ酸配列や2つのプロファイル行列中のどの残基又はプロファイルカラムを対応付させるか (ここでは残基とギャップとの対応付も含まれる) 決定する必要があるが、その対応付のさせ方は非常に多数考えられる。ダイナミックプログラミングは、その中から類似性スコアが最大となるような対応付を自動的に効率良く見出すアルゴリズムである。そしてまた、その対応付の結果それ自体が最終的に得たいアラインメントである。

ダイナミックプログラミングでは、通常のアミノ酸配列比較の場合は、比較さ

れる2つのアミノ酸配列、および、比較したい2つのアミノ酸配列の各々の残基  
 ペアに対する類似性スコア（類似の度合いを示す点数）から構成されるスコア行  
 列、プロファイル行列比較の場合は、比較される2つの代表アミノ酸配列と、比  
 較したい2つのプロファイル行列の、各々のプロファイルカラムのペアに対する  
 5 類似性スコアから構成されるスコア行列の入力を要求する。これらを入力するこ  
 とによって、ダイナミックプログラミングは、通常のアミノ酸配列比較の場合は  
 、比較されるアミノ酸配列対のアラインメントとその最終スコア（類似性スコア  
 が最大となるような最適パスを見つけることにより得られたスコア値）、プロフ  
 ァイル行列比較の場合は、比較される代表アミノ酸配列のアラインメント、およ  
 10 びその最終スコアを出力する。

したがって、ダイナミックプログラミングを使用する手法によりプロファイル  
 行列を比較するためには、比較したい2つのプロファイル行列の類似性を精度よ  
 く評価したスコア行列を作成する必要がある。

2つのプロファイル間の類似の程度を示すスコア行列を算出する方法として、  
 15 Rychlewskiらが開発した手法が知られている（Rychlewski et al. (2000),  
 9:p232-241）。これは、比較したいプロファイルカラムペア間の類似性スコアを  
 、2つのプロファイルカラムを内積したものと定義づけて算出することにより、  
 比較したい2つのプロファイル行列間のスコア行列を作成するものである。

たとえば、2つのプロファイル行列、 $X = x_1 x_2 \cdots x_p \cdots x_n$ （ただし、 $x_p$ は  
 20 アミノ酸残基位置  $p$  におけるプロファイルカラム）および  $Y = y_1 y_2 \cdots y_q \cdots y_m$   
 （ただし、 $y_q$ はアミノ酸残基位置  $q$  におけるプロファイルカラム）が与えられ  
 たとき、 $n$  行  $m$  列のスコア行列の要素である、類似性スコア  $D_{qp}$ （プロファイル  
 カラム  $x_p$  およびプロファイルカラム  $y_q$  間の類似性スコア）は、下記の式によっ  
 て与えられる。

25 【数1】

$$D_{pq} = \sum_a^j x_{pa} y_{qa}$$

[式中、 $x_{pa}$  = プロファイルカラム  $x_p$  の要素

$y_{qa}$  = プロファイルカラム  $y_q$  の要素

$j$  = プロファイルカラムの要素数 (通常 20) である。]

当該手法によれば、比較したい 2 つのプロファイルカラム間において、共にア  
 ミノ酸置換が激しくない出現残基種が非常に限られている場合には、内積した値  
 も高い数値となるため、高い類似性スコアが与えられる事になる。このように出  
 現残基種が非常に限られておりアミノ酸変異が激しくない高度に保存されている  
 残基位置は、生体内での機能的あるいは、物理化学的要請から高度に保存された  
 箇所と考えられ、生物学的にも重要な位置であると考えられている。上記手法で  
 は、このような領域はその類似性を精度良く評価することができると考えられる  
 。

しかしながら、上記手法では、こうした出現残基種が限られた位置を精度良く  
 評価することができる可能性があるものの、生物学的に重要な位置であっても、  
 モチーフ内に存在する非保存位置や、タンパク質立体構造上露出していることが  
 重要で極性のみが重大な意義を占める位置、あるいはその逆に埋没部分に位置し  
 疎水性のみが保存されている位置など、アミノ酸置換が激しく生起していてもそ  
 の置換パターンに共通性があると考えられるような領域に関して精度良く評価す  
 ることができないという問題があった。

さらに、スコア行列の各要素 (類似性スコア) の平均値は負の値である事、標  
 準偏差もほぼ一定値である事が望まれるため、類似性スコアに対して正規化処理  
 を施さなければならず、煩雑であるという問題もあった。

従って、プロファイル行列間において、保存領域のみならず、非保存領域の類  
 似性も評価できる、高精度かつ簡便な手法の開発が望まれていた。

【非特許文献 1】

Rychlewski L, Jaroszewski L, Li W, Godzik A. Protein Sci 2000 Feb;9(2):232-41

【非特許文献 2】

Altschul et al., Nucleic Acids Res. (1997) 25(17):3389-3402

【非特許文献 3】

Higgins D., Thompson J., Gibson T. Thompson J.D., Higgins D.G., Gibson



T. J. (1994). Nucleic Acids Res. 22:4673-4680

【非特許文献 4】

Needleman SB, Wunsch CD, J Mol Biol. 1970 Mar;48(3):443-53

## 5 発明の開示

本発明は、タンパク質の立体構造を予測するための、タンパク質プロファイル行列同士の類似性を評価するシステムを提供することを目的とする。

すなわち、本発明は、次のようなタンパク質プロファイル行列間の類似性評価システム、タンパク質立体構造の予測システム、コンピュータをそれらシステムとして機能させるためのプログラム、そのプログラムを記録したコンピュータ読み取り可能な記録媒体等を提供する。

(1) タンパク質の立体構造を予測するための、タンパク質プロファイル行列間の類似性を評価するシステムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の2つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段とを含むシステム。

(2) (1) 記載のシステムにより作成されたスコア行列を用いることを特徴とするタンパク質立体構造の予測システム。

(3) コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロ

ファイル行列間の類似性を評価するシステムとして機能させるためのプログラムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の2つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

を含むプログラム。

(4) 上記(3)記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。

(5) タンパク質プロファイル行列間の類似性を評価する方法であって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価方法は、以下のステップ：

(a) 入力プロファイル行列と、対象プロファイル行列の2つのプロファイル行列を用意するステップと、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出するステップと、

(c) 前記相関係数からなるスコア行列を作成するステップとを含む方法。

(6) 前記対象プロファイル行列が、立体構造が既知である複数のタンパク質に基づいて作成されるプロファイル行列であり、前記入力プロファイル行列が、立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成されるプロファイル行列である上記(5)記載の類似性評価方法。

- 5 (7) 上記(5)又は(6)で得られたスコア行列を用いることを特徴とするタンパク質立体構造の予測方法。

本発明により、タンパク質プロファイル行列間の類似性を簡便かつ精度よく評価することができる。本発明により得られたスコア行列は、タンパク質立体構造を予測するのに好適に使用される。

10

#### 図面の簡単な説明

第1図は、本発明の一実施形態において使用されるハードウェア構成を示す図である。

- 15 第2図は、本発明のプロファイル行列間類似性評価システムを含む処理手順の一例を示すフローチャートである。

第3図は、本発明のプロファイル行列間類似性評価システムにおいて、各プロファイルカラムペア毎に類似性を評価し、スコア行列を作成するステップを示す図である。

- 20 第4図は、実施例1、比較例1及び比較例2において出力された予測結果の信頼度と感度とをプロットした図である。

第5図は、実施例1及び比較例3において出力された予測結果の信頼度と感度とをプロットした図である。

#### 発明を実施するための最良の形態

- 25 以下、本発明を詳細に説明する。

##### 1. 類似度評価システム

第1図は、本発明の一実施形態において使用されるハードウェア構成を示す図である。

第1図に示すように、本発明の類似性評価システムは、CPU101、ROM102、RAM103

、入力部104、情報通信送信/受信部105、出力部106、ハードディスクドライブ(HDD)107及びCD-ROMドライブ108等を備える。

CPU101は、情報記憶手段（例えば磁氣的及び／又は光学的記録媒体）に記憶されているプログラムに従って、類似性評価システム全体を制御する。そして、入力部104などから受け取った情報を出力部106に供給する。また、ネットワーク回線109を通じて受け取った情報に基づいて評価処理を実行することもできる。入力部104は、キーボードやマウス等であり、評価処理を実行する上で必要な条件又はデータを入力するときに操作される。ROM102は、本発明の類似性評価システムの動作に必要な処理を命令するプログラム等を格納する。RAM103は、類似性評価システムにおける処理を実行する上で必要なデータを一時的に格納する。

送信／受信部105は、CPU101の命令に基づいて、ネットワーク回線109等との間で情報通信（データの送受信処理）を実行するものであり、例えばモデム、ルーター等が例示される。出力部106は、入力手段104から入力されたプロファイルデータ、その他各種条件等を、CPU101からの命令に基づいて情報表示処理する（例えば表示画面、プリンタ）。CD-ROMドライブ108は、CPU101の指示に基づいて、CD-ROMに格納されている類似性評価システムを機能させるためのプログラム又はデータ等を読み出し、例えばRAM103に格納する。CD-ROMの代わりに記録媒体として書き換え可能なCD-R、CD-RWを用いることもできる。その場合には、CD-ROMドライブ108の代わりにCD-R又はCD-RW用ドライブを設ける。また、上記媒体の他に、DVD、MOとそれらの媒体を用い、それに対応するドライブを備える構成としてもよい。

コンピュータに本発明の類似性評価システムを機能させるためのプログラムは、例えばC言語等で書くことができる。従って、このソフトウェアはWindows（登録商標）95/98/2000、Linux（登録商標）、UNIX（登録商標）等の各種オペレーティングシステムで作動させることが可能である。

第2図は、本発明のプロファイル行列間類似性評価システムを含む処理手順の一例を示すフローチャートである。

第2図に示すように、本発明にかかる類似性評価システムでは、まず、比較したい2つのプロファイル行列（入力プロファイル行列と対象プロファイル行列）を用意し、続いてそれらの類似性を評価し、必要に応じて評価結果を出力する。

以下、各処理について詳細に説明する。

(a) プロファイル行列の用意 (S 1 0)

プロファイル行列を用意するステップでは、比較したい2つのプロファイル行列が用意(抽出)される(S 1 1、S 1 2)。ここで、2つのプロファイル行列のうち、一方(対象プロファイル行列)は、立体構造が既知である複数のタンパク質に基づいて作成されたプロファイル行列(第2図中、長さm)である。他方(入力プロファイル行列)は、立体構造を予測したいタンパク質(立体構造は未知であると既知であると問わない)を含む複数のタンパク質に基づいて作成されたプロファイル行列(第2図中、長さn)であることが好ましい。

10     プロファイル行列の作成方法としては、上述した従来知られている方法を採用することができ、特に制限はない。たとえば、ある一配列を入力値として、既存のプログラムであるPSI-BLASTを用いて、配列データベースに検索をかけてマルチプルアラインメントを作成し、このマルチプルアラインメントに基づいてプロファイル行列を作成してもよい。また、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列の一群を入力値として、既存のプログラムであるCLUSTALWを用いてマルチプルアラインメントを作成し、当該マルチプルアラインメントに基づいてプロファイル行列を作成してもよい。また、予め作成されたマルチプルアラインメントを入力値とし、このマルチプルアラインメントに基づいて作成してもよい。

20     ここで、プロファイル行列は、ある代表アミノ酸配列の全配列に基づいて作成されていてもよく、また、代表配列中のモチーフ領域等、一部の領域に基づいて作成されていてもよい。また、マルチプルアラインメントを作成する際に、経験的に導出されたギャップペナルティを導入してもよい。

25     また、必要に応じて、プロファイル行列として、アミノ酸種の出現頻度を、アミノ酸種のランダム出現頻度で割った行列(PSSM: Gribskov, M., et al., (1987) Proc. Natl. Acad. Sci. USA, 84, 4355-4358)を用いてもよい。

入力プロファイル行列は、たとえば、立体構造を予測したいタンパク質を代表アミノ酸配列として、この配列に基づいて作成することができる。また、対象プロファイル行列については、たとえば、SCOP (Murzin et al., J. Mol. Biol.

247(4):536-540 (1995))やCATH(Orengo et al., Structure 5(8):1093-1108 (1997))といったタンパク質構造分類データベースから取得したタンパク質のアミノ酸配列を代表配列とし、この配列に基づいて作成することができる。こうして得られた対象プロファイル行列は、代表配列ごとに予め作成しておき、対象プロファイル行列データベースとして保持しておくことが好ましい。

(b) 相関係数の算出(プロファイル行列の比較評価) (S 20)

続いて、プロファイル行列の類似性評価ステップでは、上記のステップで用意した入力プロファイル行列の各プロファイルカラムと、対象プロファイル行列の各プロファイルカラムとの間の類似性を、各カラムペア毎に評価をする。

10 第3図は、各プロファイルカラムペア毎に類似性を評価し、スコア行列を作成するステップを模式的に示した図である。

本発明において、プロファイルカラム間の類似性は、プロファイルカラム間の相関係数を算出することによって行う。

たとえば、入力プロファイル行列を $X = x_1 x_2 \cdots x_p \cdots x_n$  (ただし、 $x_p$ はアミノ酸残基位置 $p$ におけるプロファイルカラム)とし、対象プロファイル行列を $Y = y_1 y_2 \cdots y_q \cdots y_m$  (ただし、 $y_q$ はアミノ酸残基位置 $q$ におけるプロファイルカラム)としたときに、プロファイルカラム $x_p$ および $y_q$ 間の類似性スコア $c_{q,p}$ は、下記の式によって与えられる。

【数2】

$$C_{pq} = \frac{\sum_a^j (x_{pa} - \bar{x}_p)(y_{qa} - \bar{y}_q)}{\sqrt{\sum_a^j (x_{pa} - \bar{x}_p)^2 \sum_a^j (y_{qa} - \bar{y}_q)^2}}$$

[式中、 $x_{pa}$  = プロファイルカラム  $x_p$  の要素

$y_{qa}$  = プロファイルカラム  $y_q$  の要素

$\bar{x}_p$  = プロファイルカラム  $x_p$  の平均値

$\bar{y}_q$  = プロファイルカラム  $y_q$  の平均値

$j$  = プロファイルカラムの要素数 (通常 20) である。]

本発明では、プロファイルカラム間の類似性をプロファイルカラム間の相関係数によって評価する。このため、プロファイルカラム間の相関の程度によって、類似性スコアが +1 から -1 の値をとることになる。たとえば、2つのプロファイルカラム中の要素間に相関がある場合、即ちアミノ酸置換パターンの傾向に類似性が有る場合には、相関係数は +1 に近い数値を取ることになる。また、2つのプロファイルカラムの各要素が互いにランダムな値を取っている場合、即ちアミノ酸置換パターンの傾向に相関が無い場合、相関係数は 0 になり、アミノ酸置換パターンの傾向が全く反対の場合、相関係数は -1 になり、アミノ酸置換パターンの傾向性の類似-非類似を非常に自然な形で表現する事が出来る。

したがって、本発明では、アミノ酸残基の保存性が高い保存領域のような相関が高い領域では、高い類似性スコアが得られるため、保存領域の類似性を精度よく評価することができる。

また、本発明によれば、アミノ酸残基の保存性だけではなく、内積によって類似性を評価する従来の方法 (Rychlewski et alら) では不可能であった領域に関する類似性評価、たとえば、モチーフ内に存在する非保存位置や、タンパク質立体構造上露出していることが重要で極性のみが重大な意義を占める位置、あるいはその逆に埋没部分に位置し疎水性のみが保存されている位置といった、激しいアミノ酸置換があるもののその置換パターンに共通性があると考えられる領域に

についての類似性をより精度良く評価することが可能である。

例えば、あるzinc fingerモチーフを有する2つのプロファイル行列を比較した場合を考えたとする。そのモチーフは

C-[DES]-x-C-x(3)-I

- 5 と表記される。これは、1, 4, 8番目の残基にそれぞれC, C, Iの残基が保存されており、2番目の残基では、D又はE又はSが出現し、3番目および、5, 6, 7番目の残基では保存残基が特に無いことが表されている。内積によって類似性を評価する従来の方法では、この場合、1, 2, 4, 8番目の残基位置では、高い数値を与えるが、その他の位置では低い数値しか与えない。したがって、内積によって類似性を評価する従来の方法は、モチーフの一部については類似性を評価しているものの、モチーフ全体の類似性については精度よく評価していないということになる。

- 15 しかしながら、本発明によれば、1, 2, 4, 8番目の残基位置に高い数値を与えるだけでなく、3, 5, 6, 7番目の残基位置においても、保存残基が特に無いという置換パターンの類似性を評価することが可能で、これら残基位置でも高い数値を与える。したがって、本発明によれば、モチーフ全体としてのパターン情報の全てを評価することが可能となる。

- 20 なお、本発明における類似性評価システムは、モチーフ領域に限られず、立体構造を予測したいタンパク質の配列全体に適用することができる。すなわち、ギャップペナルティを導入して得られたプロファイル行列間の類似性評価にも、好適に適用することができる。

さらに、本発明によれば、スコア行列の各要素（類似性スコア）の平均値および標準偏差がほぼ一定値をとるため、類似性スコアに対する煩雑な正規化処理を施す必要がないというメリットもある。

- 25 (c) スコア行列の作成

プロファイルカラム間の相関係数（類似性スコア）は、各プロファイルカラムの全部又は一部の組合せについて算出され、これに基づいてスコア行列が作成される。スコア行列は、類似性スコアが各プロファイルカラムの全組合せについて算出された場合は、入力プロファイル行列の長さを行とし、対象プロファイル行



列の長さを列とする行列であり、類似性スコアが各プロファイルカラムの一部の組合せについて算出された場合は、その組合せの数に応じた行と列を持つ行列となる。

第2図の例では、類似性スコアは各プロファイルカラムの全組合せについて算出されており、入力プロファイル行列の長さが $n$ 、対象プロファイル行列の長さが $m$ であることから、類似性スコアは $m \times n$ 個生成される（S 2 2）。したがって、スコア行列は $n$ 行 $m$ 列となる。スコア行列は、比較したいプロファイル行列の長さ、及び算出される類似性スコアの数に応じた行列を予め定義し（S 2 1）、定義された行列の各カラムに、各プロファイルカラム間の相関係数を入力することにより作成することができる（S 2 3）。

本発明で得られたスコア行列によって、2つのプロファイル行列の最終スコア（行列間の類似性）を精度よく算出することができる。最終スコアは既知の手法により作成することができる。たとえば、第2図の例では、比較されるプロファイル行列のそれぞれの代表アミノ酸配列と、本発明によって得られたこれらのプロファイル行列間のスコア行列を入力値として、ダイナミックプログラミングを用いて最適パスを算出する（S 2 4）ことによって最終スコアを求めることができる（S 2 5）。

以上の操作を、対象プロファイル行列データベースに保持してある対象プロファイル行列のすべてに対して行うことが好ましい。

## 20 2. タンパク質立体構造の予測（S 3 0）

対象プロファイル行列ごとに得られた最終スコアは、タンパク質立体構造を予測するのに好適に使用される。たとえば、以下の既知の手順にしたがって処理をされる。

### (1) 入力値

25 まず、予測対象配列を含む入力プロファイル行列と、立体構造が既知である代表アミノ酸配列を含む対象プロファイル行列との最終スコア、および各代表配列の長さが入力される。このとき、対象プロファイル行列データベース中に $N$ 本の既知代表配列があれば、 $N$ 個の最終スコアと配列長が入力されることになる。

### (2) 最終スコアの長さ依存性の補正

予測対象配列を含む入力プロファイル行列と、各既知代表配列を含む対象プロファイル行列との最終スコアは、代表配列長に依存した関係が認められる為、次のような統計処理を行う。まず、X軸に各代表配列の長さの自然対数をとった値、Y軸に予測対象配列を含む入力プロファイル行列と各既知代表配列を含むプロファイル行列との最終スコアをプロットし、異常なはずれ値を除いて回帰直線を引く。各長さにおける最終スコアの平均値は回帰直線で表されるものとみなし、予測対象配列を含む入力プロファイル行列と各既知代表配列を含む対象プロファイル行列との最終スコアは、平均値からのずれで評価される。通常良く使用されるように、標準偏差を単位として、そのずれの度合いが測定される。

### 10 (3) ソート

平均値からのずれが（高得点側に）大きいもの程類似性が有るとみなされる。それ故、平均値からのずれが（高得点側に）大きい順にソートされ、予測構造の候補とされる。

### (4) 予測構造としてのアラインメントとスコア出力

15 上でソートされた順に予測構造の候補として出力される。結果全てを出力するのは無意味なため、予測精度を考慮し経験的に求められた閾値以上の平均値からのずれを有する結果のみを出力する。この時、予測精度の指標として、標準偏差を単位として計算される平均値からのずれの度合いが表示される。

20 予測対象配列を含む入力プロファイル行列と、各既知代表配列を含む対象プロファイル行列とのアラインメントおよび最終スコアの結果は、ダイナミックプログラミングを用いて逐次計算された際のものを出力する。各既知代表配列は立体構造既知なので、このアラインメント出力が立体構造予測結果に相当する。

## 3. コンピュータプログラム

25 本発明は、コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムとして機能させるためのプログラムをも提供する。本発明のコンピュータプログラムは、以下の手段：

(a) 入力プロファイル行列と、対象プロファイル行列の2つのプロファイル行列を用意する手段と、

(b) 前記入カプロファイル行列の各プロファイルカラムと、前記対象プロファイ

ル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段とを含むものである。

- 5 本発明のプログラムには、上記必須の手段以外に、汎用のプログラムとして通常備えられる汎用手段を含んでもよい。そのような手段としては、各種データの格納手段、情報の送受信手段、ディスプレイ、プリンター等の表示・出力手段等を挙げることができる。

#### 4. コンピュータ用記録媒体

- 10 本発明のプログラムは、コンピュータ読み取り可能な記録媒体又はコンピュータに接続しうる記憶手段に保存することができる。本発明のプログラムを含有するコンピュータ用記録媒体又は記憶手段も本発明に含まれる。記録媒体又は記憶手段としては、磁氣的媒体（フレキシブルディスク、ハードディスクなど）、光学的媒体（CD、DVDなど）、磁気光学的媒体（MO、MD）などが挙げられる。

#### 15 【実施例】

以下、実施例により本発明をさらに具体的に説明する。但し、本発明はこれら実施例に限定されるものではない。

##### 実施例 1

##### (1) 対象プロファイル行列データベースの構築

- 20 構造分類データベース S C O P (URL: <http://scop.mrc-lmb.cam.ac.uk/scop/>) release1.59 に基づく分類から、代表配列を取得した。その中から、単独ドメインを有し解像度 2.5 Å 以内の構造データを有するタンパク質のアミノ酸配列 9 4 8 本を選択した。9 4 8 本の代表配列各々に対して PSI-BLAST とアミノ酸配列データベース (N R D B: <ftp://ftp.ncbi.nlm.nih.gov> より取得) を用いて対象プロファイル行列を構築し、対象プロファイル行列データベースを完成させた。

25 ここで使用した「N R D B」には、現在知られているほぼ大部分のタンパク質アミノ酸配列が含まれている。PSI-BLAST を使うことで、この N R D B から各代表配列に生物学的に関連あると考えられる配列を自動的に収集し、さらにプロファイル行列も作成することが出来る。

## (2) 入力プロファイル行列の作成

本発明にかかるシステムによって正しい構造予測がなされているかどうかを調べるため、予測対象配列として構造が既に知られている配列、すなわち、対象プロファイル行列を作成する際に使用した上記 948 本の代表配列を使用した。入力プロファイル行列は、これらの予測対象配列を順次使用して、対象プロファイル行列の場合と同様の操作、すなわち、PSI-BLASTとアミノ酸配列データベース (NRDB) を用いて構築した。

## (3) 各プロファイル行列間の比較

続いて、上記で構築された予測対象配列 (本実施例では 948 本の各代表配列) を含む入力プロファイル行列と、対象プロファイル行列データベース中の対象プロファイル行列との比較が順次なされた。この際、プロファイル行列間のスコア行列の各要素 (類似性スコア) は、相関係数を用いて計算された。

こうして得られたプロファイル行列間のスコア行列を入力値として、ダイナミックプログラミングによってプロファイル行列間の最終スコアとアラインメントが出力された。

各入力プロファイル行列に対して、以上の操作を対象プロファイル行列データベースに構築されたすべての対象プロファイル行列について行った。

## (4) 最終処理及び結果出力

評価の出力は、既に説明した方法に従って、948 予測について各々結果出力を行った。すなわち、入力プロファイル行列と対象プロファイル行列との各最終スコアおよび各代表配列の長さを入力し、最終スコアの長さ依存性の補正を行った。続いて、平均値からのずれが (高得点側に) 大きい順にソートし、ソートされた順に予測構造の候補として出力した。

こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第 4 図に示した。

### 比較例 1

実施例 1 で取得した 948 本の代表配列を用いて、配列類似性検索として一般的に用いられている PSI-BLAST を用いて構造予測を行った。すなわち、

9 4 8 本の代表配列各々に対してPSI-BLASTとアミノ酸配列データベース(N R D B:ftp://ftp.ncbi.nlm.nih.govより取得)を用いて構築したプロファイル行列を入力値とし、9 4 8 本の代表配列に対して類似性検索を行い、予測構造の候補を出力した。

- 5      こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第4図に示した。

#### 比較例 2

- 実施例1で取得した9 4 8 本の代表配列を用いて、配列類似性検索として一般的に用いられているIMPALA(Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L., and Altschul, S. F. (1999) Bioinformatics. 015:1000-1011)を用いて構造予測を行った。すなわち、9 4 8 本の代表配列を入力値とし、9 4 8 本の代表配列各々に対して予め作成し構築したプロファイル行列データベース(実施例1で構築した対象プロファイル行列データベースを使用  
15      した)に対して類似性検索を行い、予測構造の候補を出力した。

こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第4図に示した。

- 第4図から、比較例1および2の手法に比べて、信頼度0.98以降において  
20      、本発明にかかる実施例1が常に感度で勝っていることが示される。

#### 比較例 3

プロファイル行列間のスコア行列の各要素(類似性スコア)を、内積法(Rychlewski et al. (2000), 9:p232-241)を用いて計算した以外は実施例1と同様の手法で予測構造の候補を出力した。

- 25      こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第5図に示した。

#### 実施例 2

##### (1) 対象プロファイル行列データベースの構築

配列は、構造分類データベース SCOP (URL: <http://scop.mrc-lmb.cam.ac.uk/scop/>) release1.59に基づく分類から、お互いの同一残基率が40%未満であるドメイン単位の代表配列4381本を、SCOPの配列データベースであるASTRAL (<http://astral.stanford.edu/>) データベースから取得した。更に、タンパク質立体構造データベース PDB (URL: <http://www.rcsb.org/pdb/>) に登録されているが、SCOPに未登録であるものであって、ASTRALから取得した上記4381本の配列と非類似のものを下記 (A) ~ (D) の要領で取得し、代表配列に加えた。このようにして選択されたアミノ酸配列各々に対して、下記 (A) ~ (D) の要領で PSI-BLAST と NRDB を用いて対象プロファイル行列を構築し、対象プロファイル行列データベースを完成させた。

(A) 対象プロファイル行列データベース A の構築

2002年5月18日時点での PDB 中のアミノ酸配列を SCOP release1.59 の分類に基づく代表配列に対して BLASTP (Altschul et al., Nucleic Acids Res. (1997) 25(17): 3389-3402; 非特許文献 2) をかけ、期待値が 0.00001 以上のものを選んだ。さらにそれらを配列のクラスタリングを行うプログラムである blastclust 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 1042 1043 1044 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 1055 1056 1057 1058 1059 1060 1061 1062 1063 1064 1065 1066 1067 1068 1069 1070 1071 1072 1073 1074 1075 1076 1077 1078 1079 1080 1081 1082 1083 1084 1085 1086 1087 1088 1089 1090 1091 1092 1093 1094 1095 1096 1097 1098 1099 1100 1101 1102 1103 1104 1105 1106 1107 1108 1109 1110 1111 1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125 1126 1127 1128 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1140 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1155 1156 1157 1158 1159 1160 1161 1162 1163 1164 1165 1166 1167 1168 1169 1170 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 1181 1182 1183 1184 1185 1186 1187 1188 1189 1190 1191 1192 1193 1194 1195 1196 1197 1198 1199 1200 1201 1202 1203 1204 1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1230 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260 1261 1262 1263 1264 1265 1266 1267 1268 1269 1270 1271 1272 1273 1274 1275 1276 1277 1278 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1290 1291 1292 1293 1294 1295 1296 1297 1298 1299 1300 1301 1302 1303 1304 1305 1306 1307 1308 1309 1310 1311 1312 1313 1314 1315 1316 1317 1318 1319 1320 1321 1322 1323 1324 1325 1326 1327 1328 1329 1330 1331 1332 1333 1334 1335 1336 1337 1338 1339 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1350 1351 1352 1353 1354 1355 1356 1357 1358 1359 1360 1361 1362 1363 1364 1365 1366 1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1380 1381 1382 1383 1384 1385 1386 1387 1388 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1400 1401 1402 1403 1404 1405 1406 1407 1408 1409 1410 1411 1412 1413 1414 1415 1416 1417 1418 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1429 1430 1431 1432 1433 1434 1435 1436 1437 1438 1439 1440 1441 1442 1443 1444 1445 1446 1447 1448 1449 1450 1451 1452 1453 1454 1455 1456 1457 1458 1459 1460 1461 1462 1463 1464 1465 1466 1467 1468 1469 1470 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500 1501 1502 1503 1504 1505 1506 1507 1508 1509 1510 1511 1512 1513 1514 1515 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530 1531 1532 1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1544 1545 1546 1547 1548 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1560 1561 1562 1563 1564 1565 1566 1567 1568 1569 1570 1571 1572 1573 1574 1575 1576 1577 1578 1579 1580 1581 1582 1583 1584 1585 1586 1587 1588 1589 1590 1591 1592 1593 1594 1595 1596 1597 1598 1599 1600 1601 1602 1603 1604 1605 1606 1607 1608 1609 1610 1611 1612 1613 1614 1615 1616 1617 1618 1619 1620 1621 1622 1623 1624 1625 1626 1627 1628 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1649 1650 1651 1652 1653 1654 1655 1656 1657 1658 1659 1660 1661 1662 1663 1664 1665 1666 1667 1668 1669 1670 1671 1672 1673 1674 1675 1676 1677 1678 1679 1680 1681 1682 1683 1684 1685 1686 1687 1688 1689 1690 1691 1692 1693 1694 1695 1696 1697 1698 1699 1700 1701 1702 1703 1704 1705 1706 1707 1708 1709 1710 1711 1712 1713 1714 1715 1716 1717 1718 1719 1720 1721 1722 1723 1724 1725 1726 1727 1728 1729 1730 1731 1732 1733 1734 1735 1736 1737 1738 1739 1740 1741 1742 1743 1744 1745 1746 1747 1748 1749 1750 1751 1752 1753 1754 1755 1756 1757 1758 1759 1760 1761 1762 1763 1764 1765 1766 1767 1768 1769 1770 1771 1772 1773 1774 1775 1776 1777 1778 1779 1780 1781 1782 1783 1784 1785 1786 1787 1788 1789 1790 1791 1792 1793 1794 1795 1796 1797 1798 1799 1800 1801 1802 1803 1804 1805 1806 1807 1808 1809 1810 1811 1812 1813 1814 1815 1816 1817 1818 1819 1820 1821 1822 1823 1824 1825 1826 1827 1828 1829 1830 1831 1832 1833 1834 1835 1836 1837 1838 1839 1840 1841 1842 1843 1844 1845 1846 1847 1848 1849 1850 1851 1852 1853 1854 1855 1856 1857 1858 1859 1860 1861 1862 1863 1864 1865 1866 1867 1868 1869 1870 1871 1872 1873 1874 1875 1876 1877 1878 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051 2052 2053 2054 2055 2056 2057 2058 2059 2060 2061 2062 2063 2064 2065 2066 2067 2068 2069 2070 2071 2072 2073 2074 2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085 2086 2087 2088 2089 2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100 2101 2102 2103 2104 2105 2106 2107 2108 2109 2110 2111 2112 2113 2114 2115 2116 2117 2118 2119 2120 2121 2122 2123 2124 2125 2126 2127 2128 2129 2130 2131 2132 2133 2134 2135 2136 2137 2138 2139 2140 2141 2142 2143 2144 2145 2146 2147 2148 2149 2150 2151 2152 2153 2154 2155 2156 2157 2158 2159 2160 2161 2162 2163 2164 2165 2166 2167 2168 2169 2170 2171 2172 2173 2174 2175 2176 2177 2178 2179 2180 2181 2182 2183 2184 2185 2186 2187 2188 2189 2190 2191 2192 2193 2194 2195 2196 2197 2198 2199 2200 2201 2202 2203 2204 2205 2206 2207 2208 2209 2210 2211 2212 2213 2214 2215 2216 2217 2218 2219 2220 2221 2222 2223 2224 2225 2226 2227 2228 2229 2230 2231 2232 2233 2234 2235 2236 2237 2238 2239 2240 2241 2242 2243 2244 2245 2246 2247 2248 2249 2250 2251 2252 2253 2254 2255 2256 2257 2258 2259 2260 2261 2262 2263 2264 2265 2266 2267 2268 2269 2270 2271 2272 2273 2274 2275 2276 2277 2278 2279 2280 2281 2282 2283 2284 2285 2286 2287 2288 2289 2290 2291 2292 2293 2294 2295 2296 2297 2298 2299 2300 2301 2302 2303 2304 2305 2306 2307 2308 2309 2310 2311 2312 2313 2314 2315 2316 2317 2318 2319 2320 2321 2322 2323 2324 2325 2326 2327 2328 2329 2330 2331 2332 2333 2334 2335 2336 2337 2338 2339 2340 2341 2342 2343 2344 2345 2346 2347 2348 2349 2350 2351 2352 2353 2354 2355 2356 2357 2358 2359 2360 2361 2362 2363 2364 2365 2366 2367 2368 2369 2370 2371 2372 2373 2374 2375 2376 2377 2378 2379 2380 2381 2382 2383 2384 2385 2386 2387 2388 2389 2390 2391 2392 2393 2394 2395 2396 2397 2398 2399 2400 2401 2402 2403 2404 2405 2406 2407 2408 2409 2410 2411 2412 2413 2414 2415 2416 2417 2418 2419 2420 2421 2422 2423 2424 2425 2426 2427 2428 2429 2430 2431 2432 2433 2434 2435 2436 2437 2438 2439 2440 2441 2442 2443 2444 2445 2446 2447 2448 2449 2450 2451 2452 2453 2454 2455 2456 2457 2458 2459 2460 2461 2462 2463 2464 2465 2466 2467 2468 2469 2470 2471 2472 2473 2474 2475 2476 2477 2478 2479 2480 2481 2482 2483 2484 2485 2486 2487 2488 2489 2490 2491 2492 2493 2494 2495 2496 2497 2498 2499 2500 2501 2502 2503 2504 2505 2506 2507 2508 2509 2510 2511 2512 2513 2514 2515 2516 2517 2518 2519 2520 2521 2522 2523 2524 2525 2526 2527 2528 2529 2530 2531 2532 2533 2534 2535 2536 2537 2538 2539 2540 2541 2542 2543 2544 2545 2546 2547 2548 2549 2550 2551 2552 2553 2554 2555 2556 2557 2558 2559 2560 2561 2562 2563 2564 2565 2566 2567 2568 2569 2570 2571 2572 2573 2574 2575 2576 2577 2578 2579 2580 2581 2582 2583 2584 2585 2586 2587 2588 2589 259

2002 年 7 月 14 日時点での PDB と 2002 年 6 月 23 日時点での PDB 中のアミノ酸配列の差分を上記 (B) で作成した代表配列に対して BLASTP をかけ、期待値が 0.00001 以上のものを選んだ。さらにそれらを blastclust にかけ、互いの同一残基率が 40%未満となるように配列 23 本を選択した。このようにして選択された配列と、上記 (B) で作成した代表配列との合計 4701 本の配列各々に対して、PSI-BLAST と 2002 年 7 月 9 日時点の NRDB を用いて対象プロファイル行列を構築し、対象プロファイル行列データベース C を完成させた。

#### (D) 対象プロファイル行列データベース D の構築

上記 (C) で作成した代表配列の合計 4701 本の配列各々に対して、PSI-BLAST と 2002 年 8 月 6 日時点の NRDB を用いて対象プロファイル行列を構築し、対象プロファイル行列データベース D を完成させた。

#### (2) 入力プロファイル行列の作成

配列は、隔年で行われる世界的規模で行われる構造予測コンテストの 2002 年度大会である CASP5/CAFASP3 (URL: [http:// predictioncenter.llnl.gov/casp5/](http://predictioncenter.llnl.gov/casp5/)) において、構造認識部門 (通常の配列解析手法では立体構造既知であるタンパク質と明白な配列類似性を有さないが、その構造が (実際に解かれてみると) 既知立体構造との構造類似性を有する、即ち類似性検索が困難なタンパク質に関する予測部門) において出題された配列、すなわち、現在通常の配列解析手法 (例えば、PSI-BLAST など) では、立体構造既知であるタンパク質と明白な配列類似性を有さないタンパク質であり、かつ、その構造が (実際に解かれてみると) 既知立体構造との構造類似性が明らかになったアミノ酸配列を用いた。具体的には、URL: <http://www.cs.bgu.ac.il/~dfischer/CAFASP3/targets.html> において、下記のターゲット番号が付されたアミノ酸配列 22 本を用いた。

T0130、T0132、T0134、T0135、T0136、T0138、T0146、T0147、T0148、T0156、T0157、  
T0159、T0162、T0168、T0170、T0172、T0173、T0174、T0186、T0187、T0191、T0193

これら 22 本の配列各々に対して、PSI-BLAST と NRDB を用いて入力プロファイル行列を構築し、入力プロファイル行列データベースを完成させた。

なお、NRDB としては、2002 年 5 月 18 日時点、2002 年 6 月 17 日時点、2002 年 7 月 9 日時点、及び 2002 年 8 月 6 日時点のものの計 4 種類を使用し、得られた

入力プロファイル行列データベースを、それぞれ、「入力プロファイル行列データベースA」、「入力プロファイル行列データベースB」、「入力プロファイル行列データベースC」、及び「入力プロファイル行列データベースD」とした。

### (3) 各プロファイル行列間の比較

- 5 続いて、上記で構築された予測対象配列を含む入力プロファイル行列データベースAの入力プロファイル行列と、対象プロファイル行列データベースA中の対象プロファイル行列との比較を、実施例1の「(3)各プロファイル行列間の比較」と同様の手順で行った（比較A）。

- 同様の操作を、入力プロファイル行列データベースBと対象プロファイル行列データベースBに対して、入力プロファイル行列データベースCと対象プロファイル行列データベースCに対して、及び、入力プロファイル行列データベースDと対象プロファイル行列データベースDに対して、それぞれ行った（比較B, C, D）。

### (4) 最終処理及び結果出力

- 15 評価の出力は、既に説明した方法に従って22予測について各々結果出力を行った。即ち、各データベースの組合せ（比較A～D）においてそれぞれ得られた、入力プロファイル行列と対象プロファイル行列との各最終スコアおよび、各代表配列の長さを入力し、最終スコアの長さ依存性を補正した。続いて平均値からのずれが、（高得点側に）大きい順にソートし、ソートされた順に上位10個までを  
20 予測構造の候補として22本の配列各々に対して出力した（出力A～D）。

- こうして出力された予測構造の候補と、コンテストの予測構造投稿期間の後に公開された実験により解かれた立体構造とを比較することで、予測結果の正確さが測定された。予測構造評価方法の一つは、予測構造と正解構造の重ね合わせを行い、対応残基が3Åより短い距離にある残基数を出力A～Dについて積算すること（sum値）により行われた。22のタンパク質を構造ドメイン単位（全部で34ド  
25 メイン）で眺めた結果によれば、構造予測コンテストCASP5/CAFASP3における上記構造認識部門において22本の配列各々に対して上位1個の予測を考慮した時、本手法のsum値は「577」であり、これは、配列情報を用いた他のいかなる手法よりも優れているものであった。



また、ある閾値を設定してある入力（予測対象）配列に対する予測の成否を観測した場合でも、22本の配列各々に対して上位1個の予測を考慮した時本手法は、予測が成功したと判断される個数を出力A～Dについて積算したもの（correct 値）において、「9」と高く、配列情報を用いた他のいかなる手法よりも優れていることが示された。

## 請求の範囲

1. タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の2つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段とを含むシステム。

2. 請求の範囲第1項に記載のシステムにより作成されたスコア行列を用いることを特徴とするタンパク質立体構造の予測システム。

3. コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムとして機能させるためのプログラムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の2つのプロファイル行列を用意する手段と、

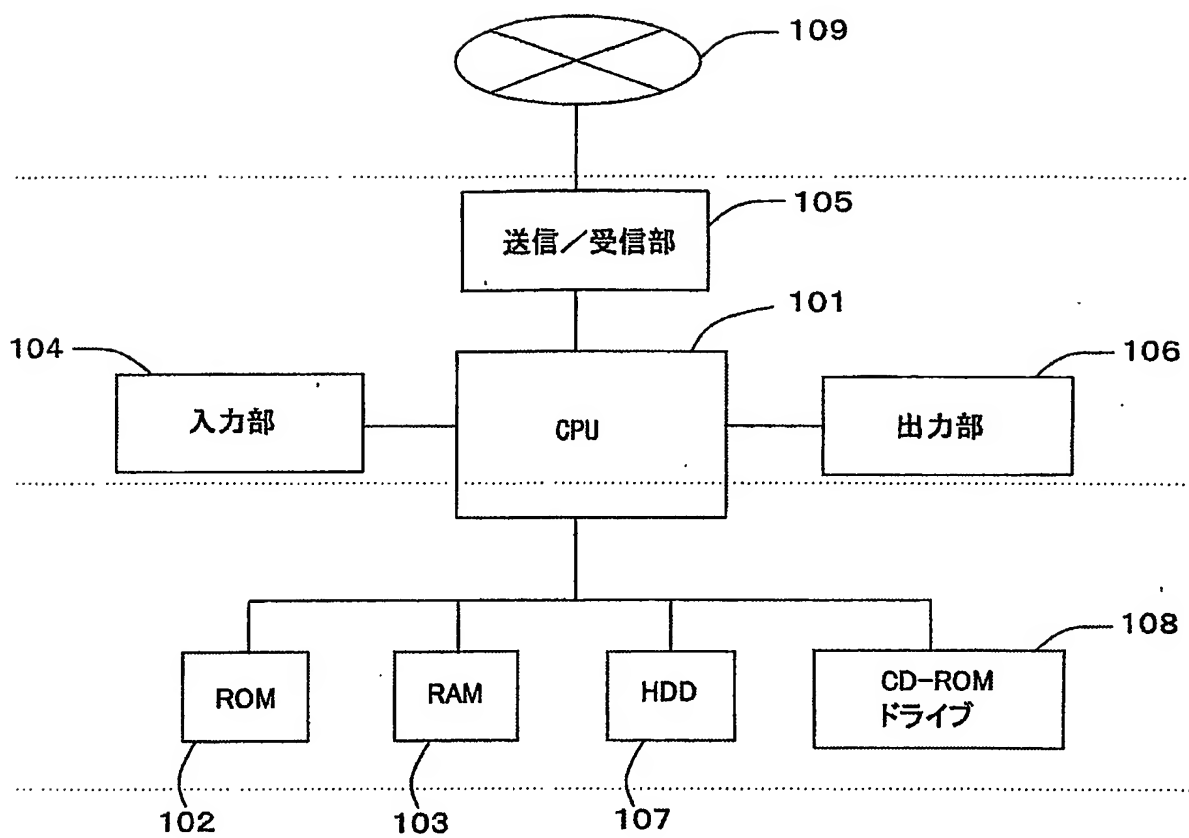
(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

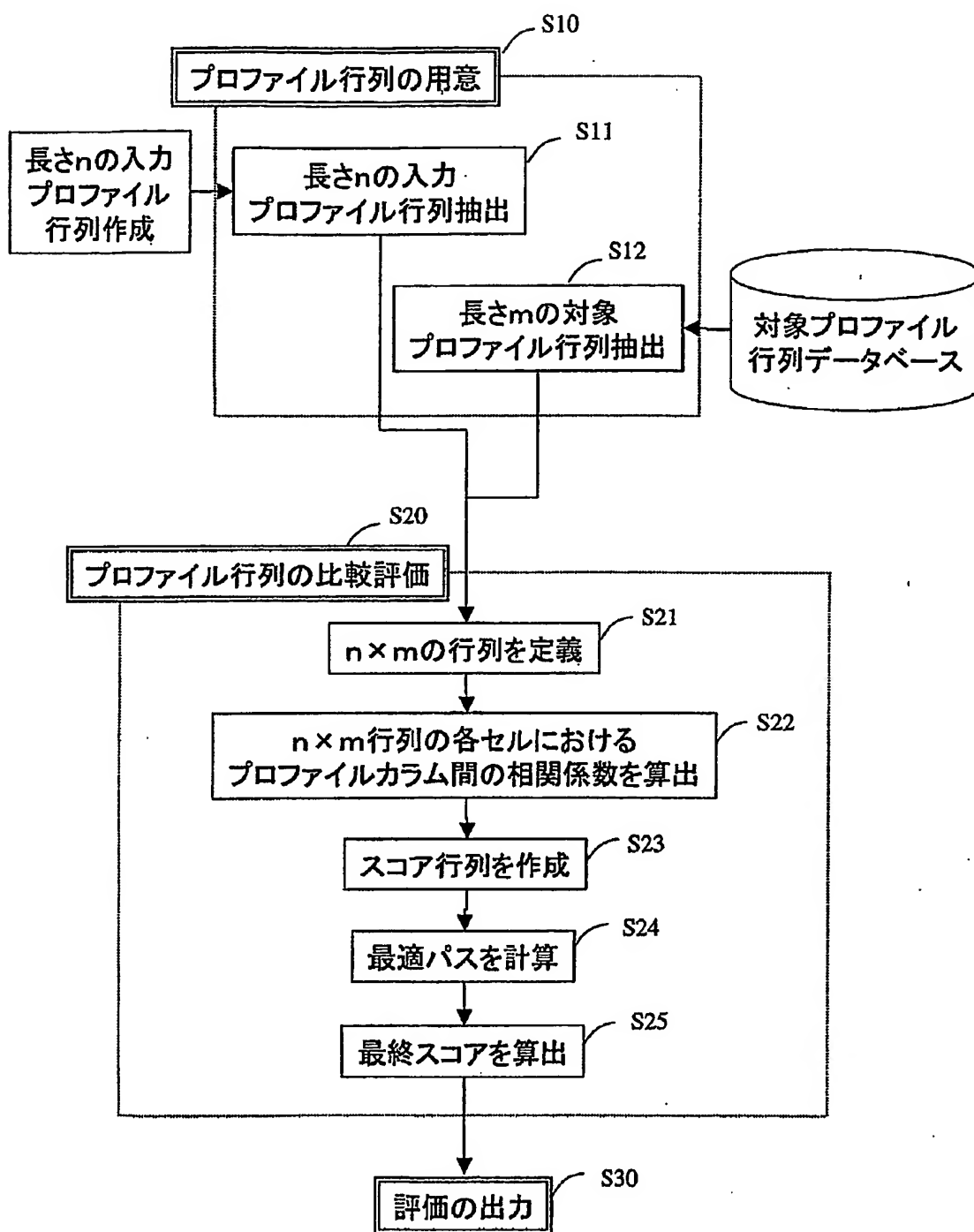
5 を含むプログラム。

4. 請求の範囲第3項に記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。

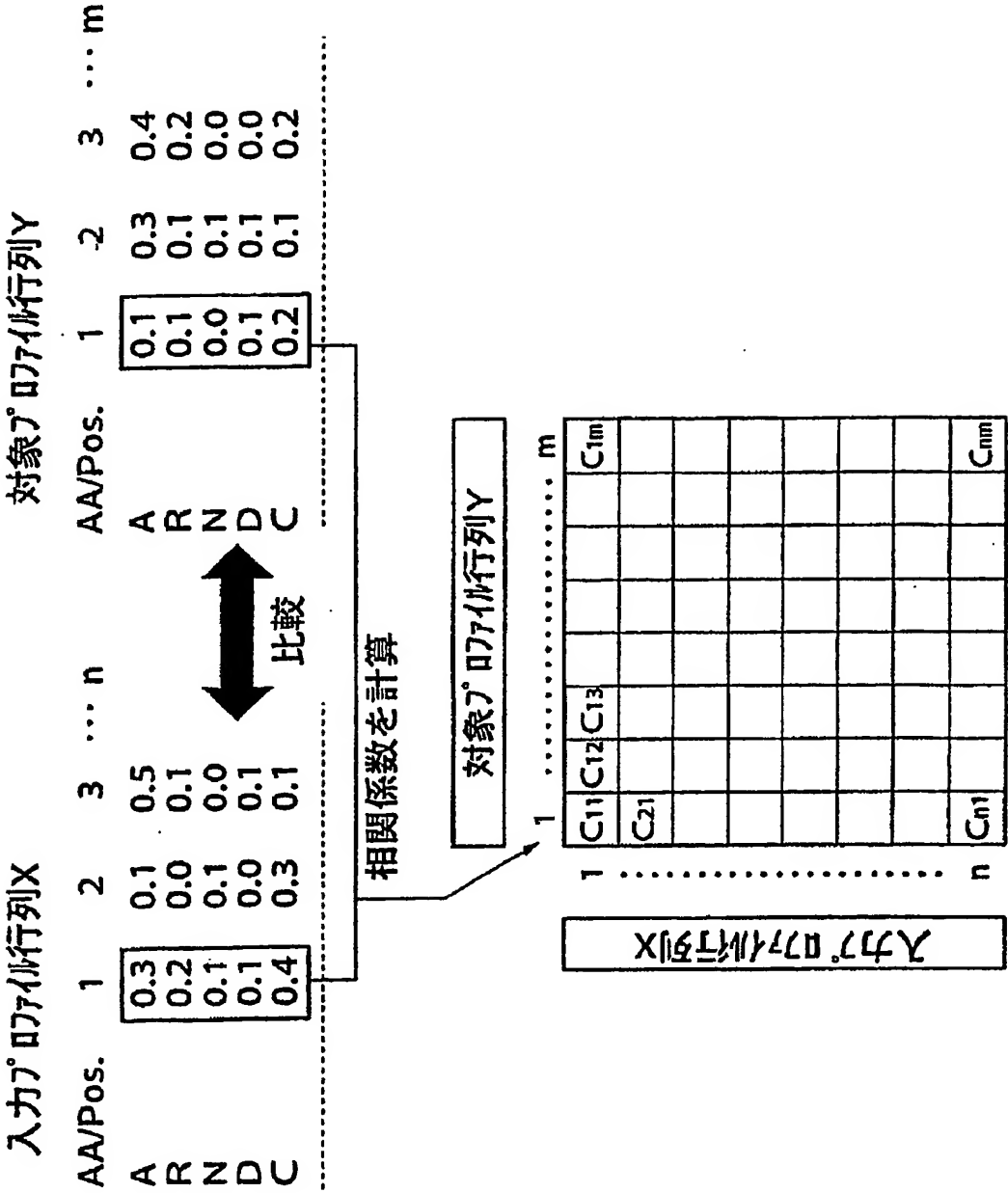
第 1 図



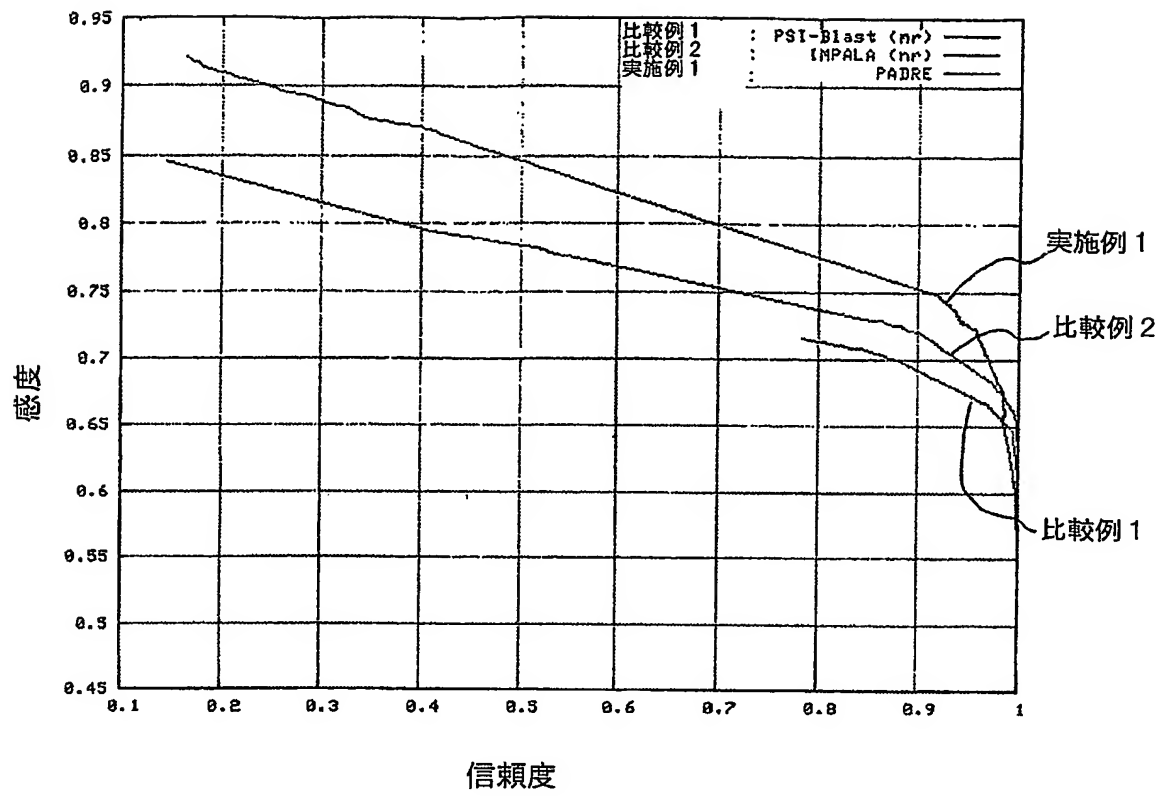
第 2 図



第 3 図



第 4 図



第 5 図

